

Tolls and Welfare Optimization for Multiclass Traffic in Multiqueue Systems

Tejas Bodas Ayalvadi Ganesh D. Manjunath
IIT Bombay, INDIA University of Bristol, UK IIT Bombay, INDIA

Abstract—We consider a queueing system with multiple heterogeneous servers serving a multiclass population. The classes are distinguished by time costs. All customers have i.i.d. service requirements. Arriving customers do not see the instantaneous queue occupancy. Arrivals are randomly routed to one of the servers and the routing probabilities are determined centrally to optimize the expected waiting cost. This is, in general, a difficult optimization problem and we obtain the structure of the routing matrix. Next we consider a system in which each queue charges an admission price. The arrivals are routed randomly to minimize an individual objective function that includes the expected waiting cost and the admission price. Once again, we obtain the structure of the equilibrium routing matrix for this case. Finally, we determine the admission prices to make the equilibrium routing probability matrix equal to a given optimal routing probability matrix.

I. INTRODUCTION

We study service systems involving customers of multiple types or classes, any of whom can be served by any one of several heterogeneous servers. Customers arrive into the system according to a random process, reside in a queue while waiting for service or being served, and then depart. Customer classes differ in their aversion to some congestion based metric like waiting time, delay, the number in the queue, etc. We seek to determine how customers may be assigned to servers in such a way as to optimize some social welfare function, and also how pricing may be used to incentivize selfish customers to achieve the same social optimum.

Examples of such systems include web server farms, cloud and grid computing clusters, communication networks and cognitive radio systems. In these examples, customers may differ in the quality of service they require, and in their willingness to pay for it. The quality of service of a customer may depend on the share of bandwidth or other resources it receives, or the service latency or the sojourn time in the system. Our modeling framework is quite general in this regard. Other examples encompassed by our framework include transport networks, which may comprise of parallel tolled and toll-free highways or multiple modes of transport, and also healthcare systems.

We capture customer requirements or preferences in the form of a congestion cost function that could be based on the occupancy, delay or waiting time in the system. We distinguish between customer classes by applying suitable multipliers to the congestion cost. Defining a social objective of minimizing the sum of customer costs, we study properties of socially optimal allocations. Next, we consider selfish customers in which each infinitesimal customer routes itself so as to optimize its

individual expected utility. In this setting, we allow each server to charge a fixed, class-independent admission price to each customer using it. The cost for each customer is the sum of the waiting time cost and the admission price. This is a non atomic routing game and we study the properties of the equilibrium distribution the traffic. Finally, we ask if there exist admission prices for which the corresponding Wardrop equilibrium coincides with the socially optimal routing; and answer it in the affirmative by determining the prices. The analysis is presented for the case when the number of classes is finite and also for the case when the number of classes is uncountable.

The end objective of this paper is optimal allocation of heterogeneous customers through admission pricing and selfish routing in a very general setting; the key generalization is that we consider arbitrary cost (equivalently, delay or any other congestion) functions at each server. In achieving this objective, we make the following key contributions.

- The optimization problem that determines the socially optimal routing policy is in general non convex and hard. However, we characterize certain structural properties possessed by *any* optimal policy. Two interesting properties of the optimal routing policy emerge in this analysis. (1) Two customer classes can have at most one common server. (2) Two servers that serve more than one traffic class between them will have distinct congestion delays.
- For the non atomic routing game with selfish customers, and arbitrary admission prices at the servers, the equilibrium distribution of traffic has a structure that is similar to the socially optimal routing.
- Finally, we show that we can indeed set prices so that the resulting equilibrium coincides with the social optimum. One set of such prices admit an interpretation as Pigouvian taxes associated with congestion externalities at the servers.

A. Previous Work

There is a substantial literature on the allocation of multi-class customers to parallel queues in both centralized and decentralized settings, including a variety of pricing schemes and game-theoretic formulations. Below, we describe some of the work more closely related to the approach taken in this paper and delineate these from the results presented in this paper. We use Kendall's notation for queueing models throughout.

Borst [1] studied the probabilistic allocation of multiclass traffic to parallel $M/G/1$ queues so as to minimize a specific social cost function, namely the total mean waiting cost per unit of time. The arriving customers are from a multiclass population and a class j is characterized by the two tuple $\left(\frac{\beta_j}{s_j}, \frac{s_j^{(2)}}{s_j}\right)$ where β_j is the waiting cost per unit time while s_j and $s_j^{(2)}$ denote the first and the second moment of the service requirement. It is shown that in an optimal allocation, customer types must be clustered according to their $\left(\frac{\beta_j}{s_j}, \frac{s_j^{(2)}}{s_j}\right)$ values and there can be at most one customer class between any two different servers. It is also shown that while each server handles a traffic mix which is as homogeneous as possible, different servers have traffic mixes of a different composition. A special case is then analyzed where the service rates at the servers are identical and when the customer classes form an ordered set. For the case when the customer classes are all identical, it is shown that load balancing will lead to an optimal allocation minimizing the total waiting cost. However when the customer classes are not identical, it is shown that an optimal allocation will never have the load at all the servers to be equal or balanced. Sethuraman and Squillante [2] considered a variant of this problem where in addition to optimal routing, one is required to also choose an order in which the multiclass customers are served in the queues. It was shown that an optimal sequencing policy for the servers is to prioritize customers of higher classes. While the routing problem could not be solved explicitly, it was shown that it had an interior solution. The structure we obtain for our optimal routing is essentially the same as in Borst [1], but our results apply to a very general class of queueing models and cost functions. Further, we also consider a game-theoretic setting of selfish optimization and a pricing mechanism that will achieve social optimality with selfish optimization. While Sethuraman and Squillante [2] also consider the optimal sequencing in a queue, we assume that the servers cannot discriminate between classes in our model, which may be more realistic depending on the application.

The equilibrium allocation of customers in a multiqueue system is studied by Bell and Stidham [3] and Haviv and Roughgarden [4]. Bell and Stidham [3] studied a single non balking traffic class served by a set of parallel $M/G/1$ queues. The servers differ in their holding cost per unit time denoted by h_i and the service rate that is denoted by μ_i for server i . With the assumption that $h_1/\mu_1 \leq h_2/\mu_2 \leq \dots \leq h_M/\mu_M$, the social cost minimization problem is solved using the generalized Lagrangian technique while the Wardrop conditions are used to identify the structure of the individually optimal routing. To compare the two allocations, it is shown that if a server is not used in an optimal allocation, then it will also not be utilized in individually optimal routing. Further, it is shown that an individually optimal routing overloads the lower indexed servers, as compared to the socially optimal one. Restricting their attention on this problem to parallel $M/M/1$ queues, Haviv and Roughgarden [4] obtain an upper bound on the price of anarchy (PoA), defined as the ratio of the total cost at the Wardrop equilibrium to that at the social optimum. In

addition to allowing a general cost function we also consider a multiclass population of customers.

There are several works that use admission prices to reduce congestion [5]–[9]. In Naor [5] and Edelson and Hilderbrand [6], customers, who belong to a single class, have to choose between paying an admission price to enter the queue, incurring a delay cost and receiving a fixed reward for service, or balking (leaving without being served). Admission prices are set by an operator who seeks to maximize revenue. If customers can observe the queue length on arrival and base their balking decision on it, then the revenue maximizing admission price exceeds the one that maximizes social welfare ([5]). However, if customers cannot observe the queue and have to base their decision on only the known arrival and service rates, then these two admission prices coincide ([6], [7]). In the latter setting, Littlechild [7] obtains the admission fee as a Pigouvian tax and shows that this will induce a socially optimal arrival rate. While Edelson and Hilderbrand [6] and Littlechild [7] consider a single class of customers and a single $M/M/1$ queue with the delay as the cost, we consider multiclass customers, multiple servers and also a general cost function. More importantly, we do not allow balking and all customers have to join one of the queues. Bradford [8] extends some of the work ([5]–[7]) to multiclass customers each with its own delay cost and reward for service and obtains the Pigouvian admission charge for each class that achieves the socially optimal allocation. The admission charge is independent of the queue from which the customer receives service. This, and the fact that the system needs to elicit information of the customer class, we believe, makes their model inapplicable in many situations. More importantly, like in the preceding references, the objective is to reduce congestion in the system by allowing customers to balk. Masuda and Whang [9] considers a similar model except that customers of different classes have the same delay cost. The model of Altman and Kameda in [10] has some similarities with that considered in this paper. However, the focus in [10] is on the uniqueness of the Nash equilibrium while we focus on the structure of the equilibrium routing. All of the above consider non atomic routing games in which each infinitesimal customer is making a selfish decision. Typically, such routing games are analyzed as congestion games (also known as exact potential games); however our model does not admit such a potential function, because different players have different delay tolerances, and hence incur different costs in the same queue.

There is also significant literature on atomic routing games, e.g., [11]–[14]. In this case, one can view each class as having a dispatcher that seeks to allocate customers of that class to servers in such a way that the expected cost for the class is minimized. Orda, Rom, and Shimkin [11] consider the atomic routing game when a number of flows selfishly split their traffic among multiple routes, and obtain equilibrium flows on the different links. Korilis, Lazar, and Orda [12] consider the allocation of bandwidths on the links such that equilibrium routing of [11] is socially optimal. Altman, Ayesta, and Prabhu [13] and Ayesta, Brun, and Prabhu [14] also study atomic routing games in non cooperative load balancing problems. For $M/G/1$ processor sharing queues, they obtain bounds on

the price of anarchy. Our interest in this paper is on non atomic games, and we look at the use of admission prices rather than capacity allocation as the mechanism to guide the system to a social optimum.

An alternative use of admission prices is in purchasing priorities [15]–[18]. This does not interest us in this paper. See Hassin and Haviv [19] for a comprehensive survey of these and other similar models. More recently, there have been papers proposing the use of differentiated prices in the Internet and studying the resultant user strategies and equilibria [20]–[23].

II. MODEL AND PROBLEM FORMULATION

We consider a system with M classes of customers and N queues. Class m customers arrive according to a Poisson process of rate λ_m , independent of other classes. The allocation of arriving customers to queues has to be made with no knowledge of current or past queue occupancies, or past arrival times or routing decisions. (It goes without saying that the future is also unknown.) Such an assumption may be less realistic for centralized allocation than when customers have to make individual decisions. Nevertheless, imposing this assumption uniformly permits clearer comparison of the two settings. Under this assumption, it is natural to restrict attention to Markovian routing policies, i.e., to policies which route customers of class i to queue j with some fixed probability p_{ij} . This is also the class of policies considered in [1], [2]. Under Markovian routing, the aggregate arrival process into queue j is a Poisson process of rate γ_j where

$$\gamma_j = \sum_{i=1}^M \lambda_i p_{ij}. \quad (1)$$

We assume that customers of all classes have the same job size distributions, and that, once they join a queue, they are treated identically within it. Associated with queue j is a cost function $D_j(\cdot)$ that specifies a cost associated with a given aggregate arrival rate. For example, the cost could be the mean sojourn time, or some higher moment of it, or the probability of the sojourn time exceeding a specified threshold. Our only assumption is that each function D_j be monotone increasing, and continuously differentiable in the interior of its domain (the set of arrival rates for which D_j is finite), with strictly positive derivative.

Finally, with each class i , we associate a positive parameter β_i that quantifies its sensitivity to delay or congestion by multiplying the cost incurred by a class i customer by β_i . The only distinction between classes is in applying different multipliers β_i to their costs in any queue. Without loss of generality, we take $\beta_1 > \beta_2 > \dots > \beta_M$; if $\beta_i = \beta_j$, we can collapse them into a single class, as customers are otherwise assumed to be identical.

The assumptions above are rather mild. We do not restrict the number of servers at a queue or the service discipline. Indeed, different queues may have different numbers of servers and employ different service disciplines. They can also employ different cost functions, for example the mean sojourn time at one queue and the second moment at another. The only requirement is that each queue treat all customers alike,

irrespective of their class. In addition to traditional queueing models, our set-up also encompasses transport models for example, where the mean journey time on a road may be some function of the traffic intensity on it. The main motivation for the assumption of Poisson arrivals is that it makes the D_j functions of a single real variable. It is not obvious how the monotonicity and differentiability assumptions would generalize if D_j were to be a function of the law of a stochastic process.

We are now ready to state the social welfare maximization problem. The objective is

$$\inf_P U(P) = \sum_{i=1}^M \sum_{j=1}^N \beta_i \lambda_i p_{ij} D_j(\gamma_j), \quad (2)$$

where the infimum is taken over all $M \times N$ right stochastic matrices $P = [[p_{ij}]]$ (defined as matrices with non-negative entries whose row sums are unity). The γ_j depend on P through (1), though this dependence has not been made explicit in the notation. Thus, the social cost is defined as the sum of the expected costs incurred by customers of different classes at different queues, weighted by the corresponding flow rates.

Next, we consider the formulation of a game between customers. Here, we allow the queues to charge admission prices, denoted by c_j at queue j . Without loss of generality, we take $c_1 > c_2 > \dots > c_N$; if $c_i = c_j$, then we can collapse these two queues into a single queue whose delay function is the inf-convolution of the delay functions of its constituent queues, i.e.,

$$D(\gamma) = \inf \{ D_i(\gamma_1) + D_j(\gamma_2) : \gamma_1, \gamma_2 \geq 0, \gamma_1 + \gamma_2 = \gamma \}.$$

The goal of a class i customer entering the system is to choose a queue j so as to minimize $c_j + \beta_i D_j(\gamma_j)$ where γ_j is determined through the strategies of all customers. We assume that the rates $\lambda_1, \dots, \lambda_M$, the cost functions $D_j(\cdot)$ and the parameters β_i , $i = 1, \dots, M$ and c_j , $j = 1, \dots, N$ are all common knowledge. Under the additional assumptions, noted earlier, that a customer does not have access to current or past queue occupancies, or the history of arrival times or decisions, its strategy is necessarily restricted to choosing a server according to a fixed probability distribution, albeit one that may depend on its class. Thus, again, the joint strategies may be represented by a right stochastic routing matrix, P . We recall the condition for such a routing matrix P to be a Wardrop equilibrium [24] is as follows.

$$\forall i, j, k \quad p_{ij} > 0 \Rightarrow c_j + \beta_i D_j(\gamma_j) \leq c_k + \beta_i D_k(\gamma_k). \quad (3)$$

The condition of (3) says that if a customer of class i has a positive probability of using a queue, then its expected cost in that queue must be no higher than its expected cost in any other queue. Further, if there is a positive probability of such a customer of using queues j and k , then the above inequality is actually an equality. Wardrop equilibrium is also the Nash equilibrium for non atomic routing games and is extensively used. See [25](Chapter 18) for an overview of non-atomic selfish routing games.

The model described in the preceding discussion considers

the case when the customers belong to one of the M classes and all customers of class i have the same parameter β_i . An alternative model to characterize the delay sensitivity of the arriving customers is by assuming that such sensitivities β are from a continuum of values. Such a model for the heterogeneous user population is more realistic where it is hard to classify customers into a fixed number of classes, e.g., those involving human customers. In this paper, in addition to discrete classes we also consider the case of a continuum of customer classes which is modeled as follows. Customer arrivals constitute a marked Poisson process of intensity $\lambda \times F$ on $\mathbb{R} \times \mathbb{R}_+$, where F is a cumulative distribution function supported on the positive reals. The marks, β , denote the delay sensitivities of the customers and are i.i.d. random variables, independent of the arrival time of the customer, and independent of the past (arrival times and marks) of the process. We assume that the marks are drawn from a distribution F which is absolutely continuous with respect to Lebesgue measure, and having a density f . We further assume that for every β in the interior of its support, $f(\beta)$ is bounded away from zero and from ∞ .

The rest of the paper is organized as follows. We describe the structure of a matrix P^* solving the welfare optimization problem (2) in Section III. In Section IV, we show that a solution P^W of (3) also has a similar structure, for any admission prices. We then show how to choose the admission prices so as to make P^W coincide with P^* , and illustrate our general results with numerical examples in Section V. We conclude in Section VI with a discussion of some open problems.

III. SOCIAL WELFARE OPTIMIZATION

Consider the optimization problem (2). As the maps $P \mapsto \gamma_j$ and D_j are continuous, the problem is one of minimizing a continuous function over the compact set of right stochastic matrices. Hence, there is a matrix P^* achieving the infimum. We will now characterize some properties of the matrix P^* below.

- Theorem 1:** • Let P^* achieve the minimum in (2) and let γ^* denote the arrival rates corresponding to P^* , as given by (1). Consider two customer classes $i_1 < i_2$, so that $\beta_{i_1} > \beta_{i_2}$. Suppose j_1 and j_2 are distinct queues such that i_1 uses j_1 and i_2 uses j_2 , i.e., $p_{i_1 j_1}^* > 0$ and $p_{i_2 j_2}^* > 0$. Then $D_{j_1}(\gamma_{j_1}^*) < D_{j_2}(\gamma_{j_2}^*)$.
- A P^* with both $p_{i_1 j_1}^*, p_{i_1 j_2}^* > 0$ and $p_{i_2 j_2}^*, p_{i_2 j_1}^* > 0$ simultaneously is not possible. In words, it is not possible to have two distinct queues, both of which are used by each of two distinct customer classes.

Proof: The proof of the first part is by contradiction. To lighten notation, we shall write β_1, λ_1 for $\beta_{i_1}, \lambda_{i_1}$. We write p_{11} for $p_{i_1 j_1}$, p_{m1} for p_{m, j_1} , D_1^* for $D_1(\gamma_{j_1}^*)$ and so on.

Suppose first that $D_1^* > D_2^*$. For sufficiently small $\epsilon > 0$, it is clear that we can find a routing matrix P such that

$$\begin{aligned} \lambda_1 p_{11} &= \lambda_1 p_{11}^* - \epsilon, & \lambda_1 p_{12} &= \lambda_1 p_{12}^* + \epsilon, \\ \lambda_2 p_{22} &= \lambda_2 p_{22}^* - \epsilon, & \lambda_2 p_{21} &= \lambda_2 p_{21}^* + \epsilon, \end{aligned}$$

and all other elements of P are the same as the corresponding elements of P^* . In words, we have shifted a quantity ϵ of the

flow of class i_1 customers from queue j_1 to queue j_2 , and an equal quantity of class i_2 customers from j_2 to j_1 , leaving all others unaffected. Clearly, the total flow rates γ under P are exactly the same as γ^* . Consequently, we can compute the change in social cost as

$$U(P) - U(P^*) = \epsilon(\beta_1 - \beta_2)(D_2^* - D_1^*).$$

Now, $\beta_1 > \beta_2$, while we assumed that $D_1^* > D_2^*$. But this implies that $U(P) < U(P^*)$, contradicting the minimality of P^* . This establishes that $D_1^* \leq D_2^*$.

Suppose next that $D_1^* = D_2^*$. Consider the routing matrix P described above, and define the matrices $P^\alpha = \alpha P + (1 - \alpha)P^*$. Then P^α is a right stochastic matrix for all $\alpha \in [0, 1]$ and it assigns non-zero amounts of class i_1 and class i_2 traffic to each of the queues j_1 and j_2 . By the same argument as above, the total flow rates γ^α induced by P^α are the same as γ^* . Moreover, $U(P^\alpha) = U(P^*)$ as P^α only differs from P^* in changing the composition of traffic at two queues j_1 and j_2 of equal cost, while keeping the total flows unchanged. Hence, P^α also achieves the minimum in (2). We now apply the Karush-Kuhn-Tucker (KKT) conditions for optimality at P^α , where p_{11}, p_{12}, p_{21} and p_{22} are all strictly between 0 and 1. (The regularity conditions needed for the KKT necessary conditions to be applicable hold because the constraints on P^* are affine.) The KKT conditions imply that

$$\frac{\partial U(P^\alpha)}{\partial p_{11}} = \frac{\partial U(P^\alpha)}{\partial p_{12}}, \quad \frac{\partial U(P^\alpha)}{\partial p_{21}} = \frac{\partial U(P^\alpha)}{\partial p_{22}}.$$

Using the definitions of U and γ_j , we can rewrite the first equality above as

$$\begin{aligned} & \beta_1 \lambda_1 D_1^\alpha + \beta_1 \lambda_1^2 p_{11}^\alpha (D_1^\alpha)' + \lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m1}^\alpha (D_1^\alpha)' \\ &= \beta_1 \lambda_1 D_2^\alpha + \beta_1 \lambda_1^2 p_{12}^\alpha (D_2^\alpha)' + \lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m2}^\alpha (D_2^\alpha)', \end{aligned}$$

where we write $(D_1^\alpha)'$ and $(D_2^\alpha)'$ to denote the derivatives of D_{j_1} and D_{j_2} evaluated at $\gamma_{j_1}^\alpha$ and $\gamma_{j_2}^\alpha$ respectively. But $D_1^* = D_2^*$ by assumption, and γ^α coincides with γ^* for all $\alpha \in [0, 1]$ by construction. Further as P^α only differs from P^* in changing the composition of traffic from Classes i_1 and i_2 at two queues j_1 and j_2 , the terms $\lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m1}^\alpha (D_1^\alpha)'$ and $\lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m2}^\alpha (D_2^\alpha)'$ are constant for any $\alpha \in (0, 1)$. Denoting these by η_1 and η_2 respectively, we obtain that

$$\beta_1 \lambda_1^2 p_{11}^\alpha (D_1^\alpha)' = \beta_1 \lambda_1^2 p_{12}^\alpha (D_2^\alpha)' + \eta_2 - \eta_1,$$

for all $\alpha \in (0, 1)$. But this is impossible because the D' are non-zero by assumption, one of p_{11}^α and p_{12}^α is an increasing function of α while the other is a decreasing function. Hence, it is also not possible for D_1^* and D_2^* to be equal.

The second part will also be proved by contradiction. Suppose that P^* satisfies $p_{i_1 j_1}^*, p_{i_1 j_2}^* > 0$ and $p_{i_2 j_2}^*, p_{i_2 j_1}^* > 0$ simultaneously. Since $p_{i_1 j_1}^* > 0$ and $p_{i_2 j_2}^* > 0$, from the first part of the theorem we have $D_{j_1}(\gamma_{j_1}^*) < D_{j_2}(\gamma_{j_2}^*)$. However from the assumption that $p_{i_2 j_1}^* > 0$ and $p_{i_1 j_2}^* > 0$, the first part of the theorem also implies $D_{j_1}(\gamma_{j_1}^*) > D_{j_2}(\gamma_{j_2}^*)$. However, $D_{j_1}(\gamma_{j_1}^*) < D_{j_2}(\gamma_{j_2}^*)$ and $D_{j_1}(\gamma_{j_1}^*) > D_{j_2}(\gamma_{j_2}^*)$ cannot be simultaneously possible and leads to a contradiction. This

means that a P^* with both $p_{i_1 j_1}^*, p_{i_1 j_2}^* > 0$ and $p_{i_2 j_2}^*, p_{i_2 j_1}^* > 0$ is not possible.

This completes the proof of the theorem. \blacksquare

We now use Theorem 1 to establish a structural property of any optimal allocation of customers to queues, i.e., of any solution of (2).

Corollary 1: Suppose P^* solves the optimization problem (2), and let γ_j^* denote the resulting flow rates, as given by (1). Consider a re-ordering of the queues such that $D_1(\gamma_1^*) \leq D_2(\gamma_2^*) \leq \dots \leq D_N(\gamma_N^*)$. Then, there exist numbers n_1, \dots, n_M , with $1 \leq n_1 \leq n_2 \leq \dots \leq n_M \leq N$, such that

$$p_{ij} \begin{cases} > 0, & \text{if } j \in \{n_{i-1} + 1, \dots, n_i - 1\}, \\ = 0, & \text{if } j \notin \{n_{i-1}, \dots, n_i\}. \end{cases}$$

Moreover, for each n_i , either p_{in_i} or $p_{i+1 n_i}$ or both are strictly positive.¹

In words, the corollary says that each customer class uses a nearly dedicated set of queues in the above ordering, with a possible overlap only at the boundaries of the sets. Note that it is possible for more than two classes of customers to use the same queue. For example, if $n_1 = n_2 = n_3 = n_4$, then customer classes 2 and 3 certainly use this queue (as they have nowhere else to go), while customer classes 1 and 4 may or may not do so. It is possible to have $n_M < N$, in which case there are some queues that aren't used by any customer class. This would be the case if, even at zero load, the delay in these queues is larger than in the alternatives.

Proof: Let P^* solve the welfare optimization problem (2). Define $Q_i = \{j : p_{ij}^* > 0\}$ to be the set of queues used by class i under P^* , and order these queues in non-decreasing order of delays. Clearly, each Q_i is non-empty. Define

$$D_i^{\max} = \max_{j \in Q_i} D_j(\gamma_j^*), \quad D_i^{\min} = \min_{j \in Q_i} D_j(\gamma_j^*).$$

By Theorem 1, $D_i^{\max} \leq D_{i+1}^{\min}$, with equality only if the same queue j attains the maximum in the first case and the minimum in the second. Moreover, in this case, every queue $k \neq j$ in Q_i has $D_k(\gamma_k^*)$ strictly smaller than $D_j(\gamma_j^*)$, while every queue $k \neq j$ in Q_{i+1} has $D_k(\gamma_k^*)$ strictly larger than $D_j(\gamma_j^*)$.

The claim of the corollary is now obvious. \blacksquare

We now turn to the question of the computational complexity of finding the optimal allocation P^* . Despite the nice structure of the optimal allocation provided by Corollary 1, it is far from obvious whether there is a polynomial time algorithm to find the optimal allocation, or even the ordering of queues at the optimal allocation. If the welfare optimization problem (2) were convex, then standard interior point methods would be able to find an optimal allocation in polynomial time. Unfortunately, this is not the case, as the following counterexample illustrates.

Consider a system with two classes of customers and 2 queues i.e., $M = N = 2$. Class i customers arrive according to a stationary Poisson process of rate λ_i while the job sizes for both classes are i.i.d. exponential with unit mean. Both servers have a unit service rate. We also assume that $\lambda_1 + \lambda_2 < 1$ and

$\beta_1 > \beta_2$. The arrivals of class i are routed to queue j with probability p_{ij} , independent of the routing of other customers. Since $p_{i1} = 1 - p_{i2}$, $U(\cdot)$ may be seen to be a function of only (p_{11}, p_{21}) taking values in $[0, 1]^2$. To simplify notation, we will use $p_1 = p_{11}$ and $p_2 = p_{21}$. The total arrival rate into the two queues is given by

$$\gamma_1 = \lambda_1 p_1 + \lambda_2 p_2, \quad \gamma_2 = \lambda_1 (1 - p_1) + \lambda_2 (1 - p_2). \quad (4)$$

Since the arrivals are Poisson, the routing is a Bernoulli sampling of the arrivals and the service time is exponential, each of the two queues is an $M/M/1$ queue with $D_j(\gamma_j) = 1/(1 - \gamma_j)$. Therefore the social cost is

$$\begin{aligned} U(p_1, p_2) &= (p_1 \lambda_1 \beta_1 + p_2 \lambda_2 \beta_2) (D_1(\gamma_1) - D_2(\gamma_2)) \\ &\quad + (\lambda_1 \beta_1 + \lambda_2 \beta_2) D_2(\gamma_2). \end{aligned} \quad (5)$$

The partial derivative of the social cost $U(p_1, p_2)$ with respect to p_1 is given by

$$\begin{aligned} \frac{\partial U(p_1, p_2)}{\partial p_1} &= \beta_1 \lambda_1 (D_1(\gamma_1) - D_2(\gamma_2)) \\ &\quad + \lambda_1 D_1'(\gamma_1) (\beta_1 \lambda_1 p_1 + \beta_2 \lambda_2 p_2) \\ &\quad - \lambda_1 D_2'(\gamma_2) (\beta_1 \lambda_1 (1 - p_1) + \beta_2 \lambda_2 (1 - p_2)). \end{aligned} \quad (6)$$

Take $p_1 = p_2 = 1/2$. Then, $\gamma_1 = \gamma_2$. As the functions D_1 and D_2 are identical, being the delays at identical $M/M/1$ queues, hence so are their derivatives, and it is easy to see that $\partial U(p_1, p_2)/\partial p_1 = 0$ at $(1/2, 1/2)$. Similarly, $\partial U(p_1, p_2)/\partial p_2$ is also zero at this point, and it follows that the social cost function has zero gradient at $(1/2, 1/2)$. If $U(\cdot, \cdot)$ were a convex function, it would follow that $(1/2, 1/2)$ is a global minimizer. We shall show that this is not the case, and hence that U cannot be a convex function.

Define $L_1 = \{(p_1, p_2) : \gamma_1 = \gamma_2\}$, which is a line segment in $[0, 1]^2$. Observe from (4) and (5) that $\gamma_1 + \gamma_2 = \lambda_1 + \lambda_2$, and that $U(p_1, p_2)$ is a constant on L_1 , with value $(\lambda_1 \beta_1 + \lambda_2 \beta_2) D_1(\frac{\lambda_1 + \lambda_2}{2})$. Now, $(1/2, 1/2) \in L_1$, so, if the global minimum of $U(p_1, p_2)$ is attained at $(1/2, 1/2)$, then it is also attained at every point on L_1 . Hence, the gradient of U , if it exists, should be zero at every point on L_1 . We show that this is not the case.

Since $\gamma_1 = \gamma_2 = (\lambda_1 + \lambda_2)/2$ on L_1 , and the functions D_1 and D_2 are identical, we see from (6) that

$$\begin{aligned} \frac{\partial U(p_1, p_2)}{\partial p_1} &= \lambda_1 D_1' \left(\frac{\lambda_1 + \lambda_2}{2} \right) \times \\ &\quad ((2p_1 - 1)\beta_1 \lambda_1 + (2p_2 - 1)\beta_2 \lambda_2). \end{aligned}$$

Now D_1' is strictly positive by assumption, so the only way that the partial derivative could be zero on L_1 is if $(2p_1 - 1)\beta_1 \lambda_1 + (2p_2 - 1)\beta_2 \lambda_2$ is identically zero on the set L_1 . But L_1 is specified as the set on which $\gamma_1 = \gamma_2$, which, by (4) implies that $(2p_1 - 1)\lambda_1 + (2p_2 - 1)\lambda_2$ is identically zero on L_1 . Since $\beta_1 \neq \beta_2$ by assumption, it is impossible for $(2p_1 - 1)\beta_1 \lambda_1 + (2p_2 - 1)\beta_2 \lambda_2$ to also be identically zero on L_1 . Thus, U cannot achieve the global minimum on all of L_1 . By contradiction, this proves that U is not a convex function.

¹We use the convention that the set $\{a, a+1, \dots, b\}$ is assumed empty if $a > b$.

A. A continuum of customer classes

The preceding results for a finite number of customer classes can be also be extended to the case when the delay sensitivities (β) are from a continuum of values. We first reformulate the social welfare maximization problem in this setting. Let $K(\cdot, \cdot)$ denote a kernel on $\mathbb{R}_+ \times \{1, 2, \dots, N\}$, where N is the total number of queues. In other words, $K(\beta, \cdot)$ is a probability distribution on $\{1, 2, \dots, N\}$ for each $\beta \in \mathbb{R}_+$, and $K(\cdot, i)$ is a Borel-measurable function on \mathbb{R}_+ for each $i \in \{1, 2, \dots, N\}$. We interpret $K(\beta, i)$ as the probability that a customer with delay sensitivity β is allocated to queue i . Thus, the class of static routing policies in the continuous setting can be identified with the set of kernels described above. Now, the welfare optimization problem is

$$\inf_K U(K) = \lambda \sum_{j=1}^N \int_{\beta=0}^{\infty} \beta K(\beta, j) D_j(\gamma_j) dF(\beta), \quad (7)$$

where $\gamma_j = \lambda \int_{\beta=0}^{\infty} K(\beta, j) dF(\beta)$.

We can characterize the optimal solution in a manner analogous to the setting with finitely many customer classes.

Theorem 2: Let K^* achieve the minimum in (7) and let γ^* denote the arrival rates corresponding to K^* . Suppose $\beta_1 > \beta_2 > 0$, and suppose i and j are distinct queues such that

$$\int_{\beta_1}^{\infty} K(\beta, i) dF(\beta) > 0 \text{ and } \int_0^{\beta_2} K(\beta, j) dF(\beta) > 0.$$

Then $D_i(\gamma_i^*) < D_j(\gamma_j^*)$.

In words, the theorem says that if queue i is used by customers with β_1 or higher and queue j is used by customers of β_2 or lower then at optimal load distribution, the cost in queue i will be less than that in queue j .

Proof: The proof is similar to that of Theorem 1 so we only sketch it briefly.

The proof is again by contradiction. Suppose first that $D_i^* > D_j^*$. We can modify K^* so as to swap a small but non-zero volume of traffic with delay sensitivity $\beta \leq \beta_2$ in queue j with an equal volume of traffic with delay sensitivity $\beta \geq \beta_1$ in queue i . As the traffic intensities at the two queues are left unchanged by this swap, so are the delays D_i^* and D_j^* . But the cost corresponding to these delays has strictly decreased for the swapped traffic (as traffic with higher delay sensitivity $\beta \geq \beta_1 > \beta_2$ has been moved to the queue with lower delay, and replaced with an equal quantity of less delay sensitive traffic), while remaining unchanged for all other traffic. Consequently, the total cost $U(\cdot)$ of the routing has been decreased. This contradicts the optimality of K^* .

Suppose next that $D_i^* = D_j^*$. Consider the swap described above, and let K^1 denote the kernel corresponding to the resulting routing. For $\alpha \in [0, 1]$, define $K^\alpha = (1-\alpha)K^* + \alpha K^1$. The volumes of traffic, γ^α , at any queue are exactly the same for every K^α , and hence so are the delays at each queue. Consequently, the total cost $U(K^\alpha)$ does not depend on α . Consequently, every K^α must be optimal.

Now consider modifying K^α by moving an ϵ quantity of traffic from queue i to queue j . Such a change causes the total

cost to increase by the quantity

$$\Delta U = \epsilon \left(D_j'(\gamma_j^*) \int_0^\infty \beta K^\alpha(\beta, j) dF(\beta) - D_i'(\gamma_i^*) \int_0^\infty \beta K^\alpha(\beta, i) dF(\beta) \right) + o(\epsilon).$$

Note in particular that, to first order, the change in cost does not depend on the composition of the traffic moved between the queues, but depends only on the externalities imposed by the move on the rest of the traffic in the queues. Now, the optimality of K^α requires that $\Delta U = o(\epsilon)$, i.e., that the expression in brackets be zero. But it is impossible that this can hold simultaneously for all $\alpha \in [0, 1]$ since $\int_0^\infty \beta K^\alpha(\beta, j) dF(\beta)$ increases with α while $\int_0^\infty \beta K^\alpha(\beta, i) dF(\beta)$ decreases with α . ■

Now define two kernels K^1 and K^2 to be equivalent if the set $\{\beta : K^1(\beta, \cdot) \neq K^2(\beta, \cdot)\}$ has F -measure zero. We now have the following corollary which characterizes the structure of any welfare maximizing allocation.

Corollary 2: Suppose K^* solves the optimization problem (7), and let γ_j^* denote the resulting flow rates. Consider a re-ordering of the queues such that $D_1(\gamma_1^*) \leq D_2(\gamma_2^*) \leq \dots \leq D_N(\gamma_N^*)$. Then, there is an $m \in \{1, 2, \dots, N\}$ and $\beta_1 > \beta_2 > \dots > \beta_m = 0$ such that K^* is equivalent to the allocation K given by $K(\beta, \cdot) = \delta_i$ for all $\beta \in [\beta_{i-1}, \beta_i)$ and $1 < i \leq m$, and $K(\beta, \cdot) = \delta_1$ for $\beta \in [\beta_1, \infty)$. Here δ_i denotes the probability distribution that puts unit mass on i .

Note that if $m = k < N$, then the servers with index greater than k are not used for allocation in the kernel K .

IV. ADMISSION PRICES AND WARDROP EQUILIBRIA

We now consider the same queueing model, but generalized to include admission prices $c_1 > c_2 > \dots > c_N$ at queues $1, 2, \dots, N$. Each customer seeks to join a queue that minimizes the sum of the admission price, which is common to all classes, and the expected delay cost, which is weighted by a class-specific sensitivity. In Section II, we modeled the resulting interaction as a game, and wrote down the conditions for a routing matrix P to be a Wardrop equilibrium in (3). We shall now show that a Wardrop equilibrium has the same structure that we demonstrated for a social optimum in the previous section.

Theorem 3: Consider two customer classes $i_1 < i_2$, so that $\beta_{i_1} > \beta_{i_2}$, and two queues $j_1 < j_2$, so that $c_{j_1} > c_{j_2}$. There is no Wardrop equilibrium P^W in which class i_1 uses queue j_2 while class i_2 simultaneously uses queue j_1 , i.e., $p_{i_1, j_2}^W > 0$ and $p_{i_2, j_1}^W > 0$.

Proof: We shall continue to use the lighter notation of the previous section.

The proof is by contradiction. Suppose such a Wardrop equilibrium exists. Since $p_{i_2, j_1}^W > 0$ and $p_{i_1, j_2}^W > 0$, we have by (3) that

$$c_2 + \beta_1 D_2^W \leq c_1 + \beta_1 D_1^W, \quad c_1 + \beta_2 D_1^W \leq c_2 + \beta_2 D_2^W.$$

Re-arranging these inequalities, we get

$$\beta_1 (D_2^W - D_1^W) \leq c_1 - c_2 \leq \beta_2 (D_2^W - D_1^W). \quad (8)$$

Since $c_1 > c_2$, the second inequality implies that $D_2^W - D_1^W$ is strictly positive. But $\beta_1 > \beta_2$, so the two inequalities together imply that $D_2^W - D_1^W \leq 0$. This is a contradiction, so such a Wardrop equilibrium cannot exist. ■

We have the following corollary, which is an analogue of Corollary 1. The proof is omitted as it is straightforward.

Corollary 3: Suppose P^W is a Wardrop equilibrium. There exist numbers n_1, \dots, n_M , with $1 \leq n_1 \leq n_2 \leq \dots \leq n_M \leq N$, such that

$$p_{ij}^W = 0, \text{ if } j \notin \{n_{i-1}, \dots, n_i\}.$$

The main difference from Corollary 1 is that we do not guarantee that all routing probabilities are strictly positive inside these ranges. Whereas, in the welfare-optimizing setting, any unused queues were necessarily those with the largest delays, now either a large delay or a high admission price or a combination of the two could result in a queue not being used by any customer class.

Following exactly the same arguments as in Theorem 3 and Corollary 3, Corollary 4 follows for the case when the delay costs of the customers are from an absolutely continuous distribution F .

Corollary 4: Suppose K^W satisfies the Wardrop equilibrium condition. There exist thresholds $\beta_1, \beta_2, \dots, \beta_N$ such that $\beta_1 \geq \beta_2 \geq \dots \geq \beta_N \in \mathbb{R}_+$ and $K^W(\beta, \cdot) = \delta_i$ for all $\beta \in (\beta_{i-1}, \beta_i)$ and $1 < i \leq N$ while $K^W(\beta, \cdot) = \delta_1$ for $\beta \in (\beta_1, \infty)$ with δ_i denoting the probability distribution that puts unit mass on i .

Since F is absolutely continuous, at each threshold β_i for $i = 1 \dots N - 1$, the following is satisfied

$$c_i + \beta_i D_i(\gamma_i) = c_{i+1} + \beta_i D_{i+1}(\gamma_{i+1}) \quad (9)$$

A natural mechanism design problem suggested by the above results is whether we can set admission prices in queues in such a way that selfish users reacting to these prices would assign themselves to queues in the proportions required for optimizing social welfare. Pigou [26] proposed the use of a charge or levy to internalize the congestion externality in transport networks, thereby guiding the system to a social optimum. Such charges are known as Pigouvian taxes and they have been studied in several contexts including queueing systems [7], [8] and transportation networks [27], [28]. However, these have typically focused on managing demand or guiding route choice, whereas here we study their use to achieve service differentiation in a multi-class setting.

Let P^* denote the routing matrix solving the social welfare optimization problem and γ^* the corresponding vector of traffic flow rates at the different queues. Now, a marginal unit of traffic at queue n increases the delay of each customer at this queue by $D'_n(\gamma_n^*)$. This imposes a cost of $\beta_m D'_n(\gamma_n^*)$ on each class m customer using this queue, of whom there are $\lambda_m p_{mn}^*$ per unit time. Thus, the total congestion externality caused by a marginal unit of traffic at queue n , which is the Pigouvian tax for this queue, is given by

$$c_n = \sum_{m=1}^M \beta_m \lambda_m p_{mn}^* D'_n(\gamma_n^*). \quad (10)$$

We shall show that, if the admission price at each queue is set equal to the Pigouvian tax at that queue, then the optimal allocation P^* is a Wardrop equilibrium.

Theorem 4: Let P^* be a routing matrix solving the social welfare optimization problem. Assume that the resulting flows γ^* are such that γ_n^* is in the interior of the domain of $D_n(\cdot)$ for each queue, n . Let the admission prices c_1, c_2, \dots, c_N at queues $1, 2, \dots, N$ be set according to (10). Then P^* is a Wardrop equilibrium of the resulting game.

Proof: Since P^* solves the constrained optimization problem in (2), the constraints on P are affine, and U is continuously differentiable at P^* , it follows that P^* must satisfy the KKT conditions: these state that there exist $A \in \mathbb{R}^{M \times N}$ and $\mathbf{b} \in \mathbb{R}^M$ such that

$$\left. \frac{\partial U(P)}{\partial p_{in}} \right|_{P=P^*} = a_{in} + b_i, \quad a_{in} \geq 0, \quad a_{in} = 0 \text{ if } p_{in}^* > 0. \quad (11)$$

Here, a_{in} is the Lagrange multiplier on the constraint $p_{in} \geq 0$, and is zero if the constraint is slack, and b_i is the Lagrange multiplier on the constraint $\sum_{m=1}^N p_{im} = 1$.

If class i uses queue n in the optimal allocation P^* , then (11) implies that, for all $m = 1, \dots, N$, we have

$$\left. \frac{\partial U(P)}{\partial p_{in}} \right|_{P=P^*} \leq \left. \frac{\partial U(P)}{\partial p_{im}} \right|_{P=P^*}.$$

Differentiating $U(P)$ defined in (2), we can rewrite this as

$$\begin{aligned} & \beta_i \lambda_i D_n(\gamma_n^*) + \sum_{j=1}^M \beta_j \lambda_j p_{jn}^* D'(\gamma_n^*) \lambda_i \\ & \leq \beta_i \lambda_i D_m(\gamma_m^*) + \sum_{j=1}^M \beta_j \lambda_j p_{jm}^* D'(\gamma_m^*) \lambda_i. \end{aligned}$$

Substituting (10) in the above, we get

$$\beta_i D_n(\gamma_n^*) + c_n \leq \beta_i D_m(\gamma_m^*) + c_m,$$

for all $m = 1, \dots, N$. This holds for every i and n such that class i uses queue n . Comparing this with (3), we see that this is exactly the condition for P^* to be a Wardrop equilibrium. This completes the proof of the theorem. ■

When the β of the arriving customer is a i.i.d. random from an absolutely continuous distribution, then the Pigouvian price at queue n is

$$c_n = \int_0^\infty \beta K^*(\beta, n) D'_n(\gamma_n^*) dF(\beta) \quad (12)$$

Setting this as the admission price to the queues and arguing as above we show below that the optimum $K^*(\cdot, \cdot)$ corresponds to a Wardrop equilibrium.

Theorem 5: Let K^* be a kernel solving the social welfare optimization problem. Let the admission prices c_1, c_2, \dots, c_N at queues $1, 2, \dots, N$ be set according to (12). Then K^* is a Wardrop equilibrium of the resulting game.

Proof: Since K^* solves the optimization problem in (7), modifying K^* by moving a non zero volume of traffic from queue i to queue j must not cause the total cost to decrease. Let β_j , for $j = 1, \dots, \beta_m$ be as in Corollary 2. Specifically, consider β^1 and ϵ such that $\beta_{j-1} < \beta^1 < \beta^1 + \epsilon < \beta_j$.

Modify K^* to K_ϵ^* by reallocating customers with delay cost $\beta \in [\beta^1, \beta^1 + \epsilon]$ to server i . This increases the arrival rate to server i by

$$h(\epsilon) := \int_{\beta^1}^{\beta^1 + \epsilon} \lambda dF(\beta).$$

The change in the social utility due to this reallocation is

$$\begin{aligned} \Delta U &= D_i(\gamma_i + h(\epsilon)) \int_0^\infty \lambda \beta K_\epsilon^*(\beta, j) dF(\beta) \\ &+ D_j(\gamma_j - h(\epsilon)) \int_0^\infty \lambda \beta K_\epsilon^*(\beta, j) dF(\beta) \\ &- D_i(\gamma_i) \int_0^\infty \lambda \beta K^*(\beta, i) dF(\beta) \\ &- D_j(\gamma_j) \int_0^\infty \lambda \beta K^*(\beta, j) dF(\beta) \\ &= h(\epsilon) \frac{(D_j(\gamma_j - h(\epsilon)) - D_j(\gamma_j))}{h(\epsilon)} \int_0^\infty \lambda \beta K^*(\beta, j) dF(\beta) \\ &- D_j(\gamma_j - h(\epsilon)) \int_{\beta^1}^{\beta^1 + \epsilon} \lambda \beta dF(\beta) \\ &+ h(\epsilon) \frac{(D_i(\gamma_i + h(\epsilon)) - D_i(\gamma_i))}{h(\epsilon)} \int_0^\infty \lambda \beta K^*(\beta, i) dF(\beta) \\ &+ D_i(\gamma_i + h(\epsilon)) \int_{\beta^1}^{\beta^1 + \epsilon} \lambda \beta dF(\beta). \end{aligned}$$

As $\epsilon \rightarrow 0$,

$$\begin{aligned} h(\epsilon) &\rightarrow 0 \\ \frac{h(\epsilon)}{\epsilon} &\rightarrow f(\beta^1) \\ \frac{1}{\epsilon} \int_{\beta^1}^{\beta^1 + \epsilon} \beta dF(\beta) &\rightarrow \beta^1 f(\beta^1). \end{aligned}$$

Dividing throughout by ϵ and taking limits as $\epsilon \rightarrow 0$, the limiting value of ΔU is

$$\begin{aligned} &\lambda f(\beta^1) D_i'(\gamma_i^*) \int_0^\infty \beta K^*(\beta, i) dF(\beta) + \lambda D_i(\gamma_i) \beta^1 f(\beta^1) \\ &- \lambda f(\beta^1) D_j'(\gamma_j^*) \int_0^\infty \beta K^*(\beta, j) dF(\beta) - \lambda D_j(\gamma_j) \beta^1 f(\beta^1). \end{aligned}$$

Optimality of K^* requires that this be non negative. Hence dividing throughout by $\lambda f(\beta^1)$ and using (12), we have

$$\beta^1 D_j(\gamma_j) + c_j \leq \beta^1 D_i(\gamma_i) + c_i.$$

The choice of β^1 in the interval (β_{j-1}, β_j) was arbitrary. Further, the choice of i was also arbitrary. Hence for all $\beta \in (\beta_{j-1}, \beta_j)$, we have

$$\beta D_j(\gamma_j) + c_j \leq \beta D_i(\gamma_i) + c_i.$$

for all $i = 1, \dots, N$.

This is the Wardrop condition and hence completes the proof. \blacksquare

V. EXAMPLES

In [29], [30], we analyze a system with two customer classes and two exponential servers with the slower server

being free and the faster server charging an admission price of c . We identify the transition points between five feasible regimes of traffic distribution as c is increased from 0 to $\beta_1/(\mu_2 - \lambda_1 - \lambda_2) - \beta_1/\mu_1$. We describe a mechanism to determine the socially optimal admission price c . All else being fixed, we also identify the admission price c that maximizes the total revenue. Finally, we present a numerical evaluation of some bounds on the price of anarchy. An analytical treatment of anything more general than this two-class two-server system would be too complex to gain any insights. Therefore, in this section we first illustrate our results with some numerical examples. We then explain some example scenarios where the assumptions may be generalized.

The first example is of a system serving five customer class with five servers i.e., $M = N = 5$ with mean delay as the cost function. Customers of class i arrive according to a stationary unit rate Poisson process and have $\beta_i = (6 - i)$. All classes have exponentially distributed job sizes with unit mean. Server j has service rate μ_j with $\mu_1 = 2$, $\mu_2 = 3$, $\mu_3 = 2.5$, $\mu_4 = 1.1$ and $\mu_5 = 1.5$. Thus the five queues are all $M/M/1$ and the mean delay at queue j is $D_j(\gamma_j) = \frac{1}{\mu_j - \gamma_j}$. P^* that minimizes $U(P)$ is obtained numerically and is as follows.

$$P^* = \begin{matrix} & Q_2 & Q_3 & Q_1 & Q_5 & Q_4 \\ \begin{matrix} Cl\ 1 \\ Cl\ 2 \\ Cl\ 3 \\ Cl\ 4 \\ Cl\ 5 \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.528 & 0.472 & 0 & 0 & 0 \\ 0 & 0.788 & 0.212 & 0 & 0 \\ 0 & 0 & 0.786 & 0.214 & 0 \\ 0 & 0 & 0 & 0.517 & 0.483 \end{pmatrix} \end{matrix}$$

This corresponds to $\gamma_1 = 0.998$, $\gamma_2 = 1.528$, $\gamma_3 = 1.26$, $\gamma_4 = 0.483$, and $\gamma_5 = 0.731$ and $D_1 = 0.998$, $D_2 = 0.679$, $D_3 = 0.806$, $D_4 = 1.62$, and $D_5 = 1.3$. Observe that in P^* the servers are reordered in increasing order of the mean delays with Q_i denoting the server with service rate μ_i . We also see that this P^* satisfies Corollary 1. With the above allocation we see that $U(P^*) = 12.47$.

Next, let the admission prices for queues 1, ..., 5 be, respectively, 2.57, 1.53, 0.7, 0.42, and 0. A Wardrop equilibrium allocation at these prices will be

$$P^W = \begin{matrix} & Q_1 & Q_2 & Q_3 & Q_4 & Q_5 \\ \begin{matrix} Cl\ 1 \\ Cl\ 2 \\ Cl\ 3 \\ Cl\ 4 \\ Cl\ 5 \end{matrix} & \begin{pmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix} \end{matrix}.$$

Clearly, P^W satisfies Corollary 3. The Pigouvian admission prices are 3.28, 2.77, 2.194, 1.59, and 1.27. It can be verified that for these prices, the Wardrop equilibrium condition is satisfied by the social welfare maximizing allocation, P^* .

Next we consider an alternative cost function with $D_n(\gamma_n)$ denoting the probability that the waiting time experienced by an arriving customer in queue n is more than a fixed amount T . Since the queues are $M/M/1$, We now have

$$D_n(\gamma_n) = \rho_n \exp((\gamma_n - \mu_n)T)$$

where $\rho_n = \gamma_n/\mu_n$. For $T = 1$ and the other parameters as in the preceding example, the optimal routing probabilities will be

$$P^* = \begin{matrix} & Q_2 & Q_3 & Q_1 & Q_5 & Q_4 \\ \begin{matrix} Cl\ 1 \\ Cl\ 2 \\ Cl\ 3 \\ Cl\ 4 \\ Cl\ 5 \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.42 & 0.58 & 0 & 0 & 0 \\ 0 & 0.61 & 0.39 & 0 & 0 \\ 0 & 0 & 0.59 & 0.41 & 0 \\ 0 & 0 & 0 & 0.37 & 0.63 \end{pmatrix} \end{matrix}.$$

Further, the Pigouvian prices such that $P^W = P^*$ for this case are 1.5789, 0.9809, 0.5555, 0.2813, and 0.1489.

We remark here that the price vector that induces a P^W that is the same as P^* is not unique. For example, changing the Pigouvian prices of all the queues by a fixed amount will also have a P^W that is the same as P^* .

A. $M/G/1 - LIFO$ and $M/G/1 - PS$ queues

Recall that in all the preceding analyses, we have assumed that customers of all classes have the same job size distributions, and that, once they join a queue, they are treated identically within it. We now describe some systems where we can relax this assumption and allow different classes to have a different job size distributions. Throughout this example, let S_i denote the mean job size of a Class i customer with the assumption that $S_1 \geq S_2 \geq \dots \geq S_N$. We assume that the service discipline at the servers is either processor-sharing (PS) or last-come-first-serve with preemption (LCFS-PR) and the service rate of each server is unity. Now define $\hat{\beta}_i := \beta_i S_i$ and note that $\hat{\beta}_1 > \hat{\beta}_2 > \dots > \hat{\beta}_N$. The mean sojourn time of a Class i customer at server j is now given by

$$D_{ij}(\rho_j) = S_i D_j(\rho_j) \quad (13)$$

where

$$D_j(\rho_j) = \frac{1}{1 - \rho_j} \quad (14)$$

and $\rho_j = \sum_{i=1}^M p_{ij} \lambda_i S_i$. The social welfare objective now is

$$U(P) = \sum_{i=1}^M \sum_{j=1}^N \hat{\beta}_i \lambda_i p_{ij} D_j(\rho_j). \quad (15)$$

Note that replacing λ_i in (2) by $\lambda_i S_i$ gives (15) and the analysis for the structural properties of P^* and P^W remains unaffected. The theorems and corollaries discussed in the preceding sections should be suitably modified with $\hat{\beta}_i$ now assuming the role of β_i .

VI. SUMMARY AND DISCUSSION

We considered a very general model of multiple parallel queues serving a heterogeneous customer population, and studied the problem of routing customers to queues so as to maximize social welfare. We characterized certain structural properties of the welfare-optimizing allocation. Next, we considered selfish routing decisions made by individual customers and characterized the structure of Wardrop equilibria. We then showed that, if queues charge admission prices, and these

are set equal to the congestion externalities at the optimal allocation, then the resulting Wardrop equilibrium coincides with the social optimum.

In the formulation of the optimization problem in (2), we do not restrict ourselves to average values of the delay or sojourn time, but allow our cost function to be an arbitrary function of these quantities. Thus the objective function could be a measure of fair allocation or just include a ‘fairness term’ in the objective function. Such a fairness term can be incorporated by choosing a suitable convex function of the chosen performance variable as the cost. By Jensen’s inequality, more unfair allocations are more heavily penalized by the expectation of the resulting cost function. Admittedly, this is somewhat implicit, and we have not calculated any particular fairness index such as the Jain index.

Our results raise a number of questions for future research. Firstly, as mentioned in Section III, the computational complexity of determining the optimal allocation is unknown; we only showed that the optimization problem is non-convex. Likewise, the computational complexity of determining the Wardrop equilibria is also unknown.

A second question concerns the informational constraints on the model. We have assumed that the parameters λ_i and β_i are known, and available as input to determining the socially optimal allocation or setting admission prices. In practice, this information is unlikely to be available, but needs to be inferred from observation. If customer classes are known upon arrival, then the arrival rates λ can easily be measured, but eliciting β truthfully can still be a challenge. The problem is much harder if customers are heterogeneous but there are either no clearly defined classes, or that class membership is unobservable, as is often likely to be the case. In such a situation, is it still possible to set admission prices in such a way as to ensure that the Wardrop equilibrium either coincides with the welfare optimizing allocation, or approximates it to within some factor?

Finally, we have assumed that a benevolent mechanism designer sets admission prices to maximize social welfare; it is interesting to ask what happens if the admission prices are set by a revenue maximizing service provider. Further, in such a revenue maximizing scenario it would be interesting to see if competing service providers can sustain differentiated services.

REFERENCES

- [1] S. C. Borst, “Optimal probabilistic allocation of customer types to servers,” in *Proceedings of ACM SIGMETRICS*, September 1995, pp. 116–125.
- [2] J. Sethuraman and M. Squillante, “Optimal stochastic scheduling in multiclass parallel queues,” in *Proceedings of ACM SIGMETRICS*, May 1999, pp. 93–102.
- [3] C. H. Bell and S. Stidham, “Individual versus social optimization in the allocation of customers to alternative servers,” *Management Science*, vol. 29, pp. 831–839, July 1983.
- [4] M. Haviv and T. Roughgarden, “The price of anarchy in an exponential multi-server,” *Operation Research Letters*, vol. 35, pp. 421–426, 2007.
- [5] P. Naor, “The regulation of queue size by levying tolls,” *Econometrica*, vol. 37, pp. 15–24, 1969.
- [6] N. Edelson and D. Hilderbrand, “Congestion toll for Poisson queueing processes,” *Econometrica*, vol. 43, pp. 81–92, 1975.
- [7] S. Littlechild, “Optimal arrival rate in a simple queueing system,” *International Journal of Production Research*, vol. 12, pp. 371–397, 1974.

- [8] R. M. Bradford, "Incentive compatible pricing and routing policies in multiserver queues," *European Journal of Operational Research*, vol. 89, pp. 226–236, 1996.
- [9] Y. Masuda and S. Whang, "Dynamic pricing for network service: Equilibrium and stability," *Management Science*, vol. 45, pp. 857–869, June 1999.
- [10] E. Altman and H. Kameda, "Equilibria for multiclass routing in multi-agent networks," in *Proceedings of the IEEE Conference on Decision and Control*, 2001.
- [11] A. Orda, R. Rom, and N. Shimkin, "Competitive routing in multi-user communication networks," *IEEE/ACM Transactions on Networking*, pp. 510–521, 1993.
- [12] Y. Korilis, A. Lazar, and A. Orda, "Capacity allocation under noncooperative routing," *IEEE Transactions on Automatic Control*, vol. 43, no. 3, pp. 309–325, 1997.
- [13] E. Altman, U. Ayesta, and B. Prabhu, "Load balancing in processor sharing systems," in *Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools (ValueTools)*, 2008, pp. 1–10.
- [14] U. Ayesta, O. Brun, and B. Prabhu, "Price of anarchy in non-cooperative load balancing," in *Proceedings of the IEEE INFOCOM*, 2010, pp. 436–440.
- [15] Balachandran, "Purchasing priorities in queues," *Management Science*, vol. 18, no. 5, pp. 319–326, January 1972.
- [16] F. Lui, "An equilibrium model of bribery," *Political Economy*, vol. 93, pp. 760–781, 1985.
- [17] S. Rao and E. Petersen, "Optimal pricing of priority services," *Operations Research*, vol. 46, no. 1, pp. 46–56, 1998.
- [18] P. Afeche and H. Mendelson, "Pricing and priority auctions in queueing systems with a generalized delay cost structure," *Management Science*, vol. 50, no. 7, July 2004.
- [19] R. Hassin and M. Haviv, *To Queue or Not to Queue*, Kluwer Academic Publishers, 2003.
- [20] A. Odlyzko, "Paris Metro pricing for the Internet," in *Proceedings of the 1st ACM Conference on Electronic Commerce*, 1999, pp. 140–147.
- [21] R. Jain, T. Mullen, and R. Hausman, "Analysis of Paris Metro pricing for QoS with a single service provider," in *Proceedings of International Workshop on Quality of Service (IWQoS)*, June 2001, Lecture Notes in Computer Science, 2001, Volume 2092/2001, pp. 44–58.
- [22] P. Dube, V.S. Borkar, and D. Manjunath, "Differential join prices for parallel queues: Social optimality, dynamic pricing algorithms and application to Internet pricing," in *Proceedings of IEEE INFOCOM*, 2002, pp. 276–283.
- [23] V. S. Borkar and D. Manjunath, "Charge-based control of DiffServ-like queues," *Automatica*, vol. 40, pp. 2043–2057, 2004.
- [24] J. G. Wardrop, "Some theoretical aspects of road traffic research communication networks," *Proceedings of Industrial and Civil Engineering*, vol. 1, pp. 325–378, 1952.
- [25] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, *Algorithmic Game Theory*, Cambridge University Press, 2007.
- [26] A. C. Pigou, *The Economics of Welfare*, Macmillan London, 1920.
- [27] M. J. Smith, "The marginal cost taxation of a transportation network," *Transportation Research: Series B*, vol. 13B, pp. 237–242, 1979.
- [28] H. Yang and H.-J. Huang, "Principle of marginal cost pricing: How does it work in a general road network?," *Transportation Research: Series A*, vol. 32, no. 1, pp. 45–54, 1998.
- [29] T. Bodas, A. Ganesh, and D. Manjunath, "Load balancing and routing games with admission price," in *Proceedings of the IEEE Conference on Decision and Control*, 2011.
- [30] T. Bodas and D. Manjunath, "On load balancing equilibria in multiqueue systems with multiclass traffic," in *Proceedings of NETGCOOP*, 2011.